

Joint alignment and phylogeny for large genomic data



Dr. Maria Anisimova,
Head of Applied Computational Genomics,
maria.anisimova@zhaw.ch

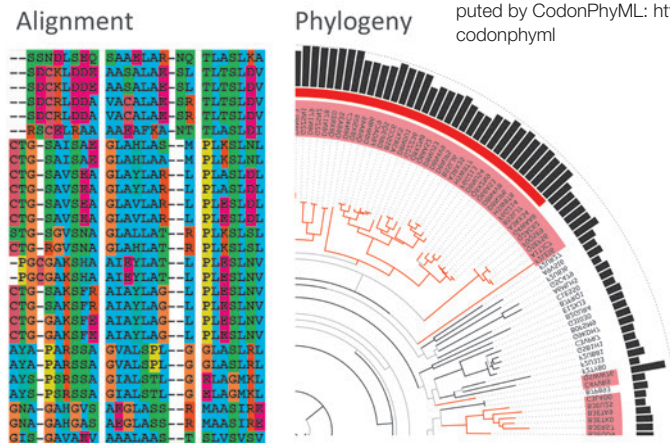
The evolutionary history of molecules is described by a tree called **phylogeny**, estimated from genomic sequences. Phylogenies are used for testing biological hypotheses with applications from medicine to ecology. Yet, phylogeny inference relies on an inferred alignment of homologous sequences, which – in turn – relies on a guide-tree reflecting their ancestral relationships. The aim is to address this apparent circularity so to improve the reliability of phylogenetic analyses.

With the advent of new generation sequencing (NGS) bioinformatic methods must keep pace to provide robust scalable solutions to analyse large data. Usually the phylogenetic inference is simplified into two independent steps: alignment inference and tree inference (Fig. 1). Since the two steps are interdependent, errors committed at each step affect the reliability of the other, and are propagated to downstream analyses. Ideally alignment and tree should be inferred jointly. Existing joint alignment-tree inference (JATI) algorithms use the Bayesian paradigm and rely on the classic evolutionary model of sequence changes based on an infinite-state continuous-time birth-death process. While useful for small datasets, these methods do not scale to large modern-day data due to the exponential time complexity of the model and the need for intensive sampling of multiple parameters including unconventional ones – alignment and tree.

Development of a new algorithm

To circumvent these problems, we will develop a new JATI algorithm in the maximum likelihood (ML) framework, building upon our methods implemented in independent packages: Codon

PhyML for fast ML phylogeny inference for protein-coding genes, and ProGraphMSA for fast probabilistic graph-based alignment. Using the Poisson process to model indels will help to reduce the time complexity to linear. The arsenal of models in CodonPhyML (Fig. 2) will improve accuracy, eg, by describing the structure of the genetic code and selection for the protein-coding genes. Further, ProGraphMSA provides one of the fastest alignment heuristics, accounts for sequence divergence, correctly penalizes insertions and deletions, and is the only alignment method that includes sequence content heterogeneity, alternative splicing and repeats.



Traditionally: Genomic Data → Alignment → Phylogeny → Hypothesis testing
Our approach: Genomic Data → (Alignment + Phylogeny) → Hypothesis testing

Fig. 1: The dependency between alignment and phylogeny estimation calls for their simultaneous inference

Analysing huge genomic datasets

Combining these features in one algorithm will result in vastly more efficient heuristics than the currently available to search the alignment-tree space. The new methodology will also allow

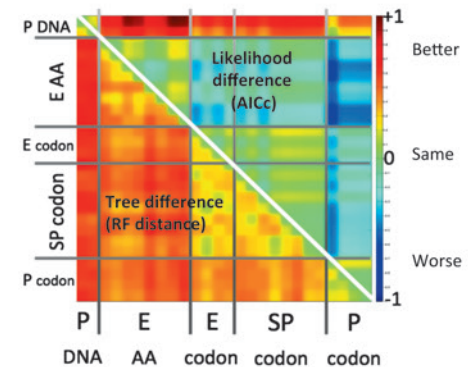


Fig. 2: A colour-coded matrix showing comparisons of 28 substitution models and correspondent phylogenies for a one-gene multiple sequence alignment, as computed by CodonPhyML: <http://sourceforge.net/projects/codonphyml>

the estimation of statistical support of inferred tree partitions and the ancestral reconstruction of molecular history. Our approach will enable analyses of huge genomic datasets. For example, we have recently applied phylogenetic methods to study large NGS datasets of maturing antibody sequences. The work was in collaboration with industry (MAB Discovery, Germany); the report is now in press in Mol Biol Evol. We expect that the new JATI method will be in demand not only in academic projects but also in pharma and biotech industry.

Research project

Fast joint estimation of alignment and phylogeny from genomic sequences in a frequentist framework

| | |
|-----------------|-----------------------------------|
| Lead: | Dr. Maria Anisimova |
| Duration: | 3 years |
| Partner: | University of Zurich |
| Funding: | Swiss National Science Foundation |
| Project budget: | CHF 454 000 |